Quality Control in State Assessments

Susan M. Brookhart, Ph.D.
Brookhart Enterprises, LLC
susanbrookhart@bresan.net
Montana Assessment Conference
April 24, 2007

Session Outline

- Overview of the standards for quality control in large-scale assessment
  - How should the meaning, accuracy, and usefulness of the information that comes from state assessment programs be warranted?
  - What are some common ways this is done in practice?

Session Outline

- Current practices of quality control used in Montana
  - How is the meaning, accuracy, and usefulness of the information that comes from Montana's state assessment program warranted?
  - What is a Technical Manual?
  - What is a Technical Advisory Committee?

Quality Control

Quality control in assessment means

- using appropriate development, administration, scoring, and reporting procedures and
- collecting and reporting evidence to document that assessment results are meaningful, accurate, and useful for intended purposes.

Interpretive argument

You might be tempted to think about testing as a "numbers game."

Validity is really more about the "**interpretive argument**" – in the same sense your English teacher would use for a theme:

- offering evidence that the **inferences** to be made from the test scores are valid,
- and the **uses** to which that information is put are valid.

Interpretive argument

- ***Scoring*** inference – assigning a score to each student's performance
- ***Generalization*** inference – generalize from the performances actually observed to the "universe of generalization" (all other similar test-like tasks under similar circumstances)
- ***Extrapolation*** inference – generalize from the universe of generalization to the broader "target domain" (trait)

Interpretive argument

- ***Implication*** inference – extend the interpretation to claims or suggestions that might be associated with verbal descriptions of the test score (e.g., "good reader")
- ***Decision*** inference – link the test scores to any decisions or actions and potential intended or unintended consequences
- ***Theory-based*** inference – extend interpretations to underlying mechanisms that account for observed performance

Interpretive argument

- ***Technical*** inference – appropriateness of assumptions regarding technical issues like
  - Equating forms
  - Scaling
  - Fit of statistical models

(Kane, 1992)

Validity

- "the degree to which evidence and theory support the **interpretations** of test scores entailed by proposed **uses** of tests."
  - First, specify intended purpose(s) and/or use(s) of the test.
  - Then, bring evidence that the relevant interpretations are warranted.

Validity evidence can be

- Based on **test content**
- Based on **response processes**
- Based on **internal structure**
- Based on relation to **other variables**
- Based on the **consequences** of testing

- A combination of these is stronger than just one for most intended purposes

Reliability

- The consistency of measures over various potential sources of error
  - Time (occasion)
  - Form
  - Rater (scorer)
- Measurement error is the converse of reliability
  - High reliability = low measurement error
  - Low reliability = high measurement error

Reliability evidence

- Test-retest correlations
- Alternate forms correlations
- Internal consistency
- Generalizability coefficients
- IRT item characteristic curves
- Standard error of measurement
  - Conditional standard error of measurement

Decision consistency

- Related concept to Reliability
- Inter-rater agreement
  - Percent
  - Kappa (% agreement corrected for amount of agreement expected by chance)

Documenting evidence of quality

- Technical manuals
- Report test development, administration, scoring, and reporting procedures so they can be reviewed by the public
- Report evidence to document that assessment results are meaningful, accurate, and useful for intended purposes (that is, report evidence for validity and reliability)

Standards: #6. Supporting Documentation for Tests

6.1 – Test documents (e.g., test manuals, technical manuals, user's guides, and supplemental material) should be made available to prospective test users and other qualified persons at the time a test is published or released for use.

Standards: #6. Supporting Documentation for Tests

6.2 – Test documents should be complete, accurate, and clearly written so that the intended reader can readily understand the contents.

Standards: #6. Supporting Documentation for Tests

6.3 – The rationale for the test, recommended uses of the test, support for such uses, and information that assists in score interpretation should be documented. Where particular misuses of a test can be reasonably anticipated, cautions against such misuses should be specified.

Standards: #6. Supporting Documentation for Tests

6.4 - intended population
item pool & scale development
description of norm group, including year

6.5 - statistical analyses supporting reliability
statistical analyses supporting validity
item level information
cut scores
raw scores and derived scores
normative data
standard errors of measurement
equating procedures

NCLB *Standards & Assessments* Peer Review Requirements

- Requires evidence for quality of
  - Content standards
  - Academic achievement standards
  - Statewide assessment system
  - Technical quality
  - Alignment
  - Inclusion
  - Reports

Technical Advisory Committees

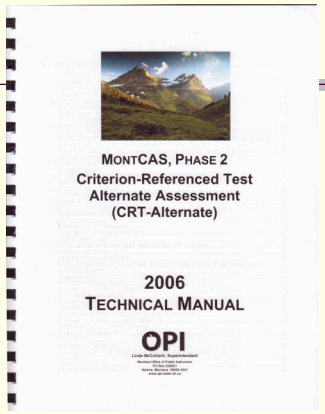
- Most states have TACs that meet at least once, and often 2 or 3 times, per year
- Committee composed of nationally recognized experts in assessment
- Usually with varying specialties
- Advice to state regarding state assessment system

Montana's Quality Control

- Technical aspects of validity documented in Technical Manuals by Measured Progress (testing contractor)
- Validity considerations about uses and consequences are the responsibility of Montana OPI
- Advice from Technical Advisory Committee

MontCAS Phase 2 CRT Tech Report

- Background & overview
- Test design
- Test development
- Design of the Reading assessment
- Design of the Mathematics assessment
- Test administration
- Scoring
- Item analyses
- Reliability
- Scaling and equating
- Reporting
- Validity summary



MontCAS Phase 2 CRT-Alt Tech Report

- Background & overview
- Overview of test design
- Test development process
- Design of the Reading assessment
- Design of the Mathematics assessment
- Test format
- Test administration
- Scoring
- Item analyses
- Reliability
- Scaling
- Reporting
- Validity summary

CRT and CRT-Alt Studies Commissioned by MT OPI

- Alignment studies
  - NWREL, 2002, 2004, 2006
- Rigor of standards study, NWREL, 2006
- CRT-Alt Inter-rater Reliability Study
  - Gail McGregor, UM, 2007
- Subgroup performance by standard
  - Art Bangert, 2003
- Independent review of technical manuals
  - Steve Sireci, 2006; Sue Brookhart, 2007
- [Studies of ITBS prior to 2003]

Montana TAC 2007

- Art Bangert, Ph.D., *Montana State University*
- Derek Briggs, Ph.D., *University of Colorado*
- Sue Brookhart, Ph.D., *Brookhart Enterprises LLC*
- Ellen Forte, Ph.D., *edCount LLC*
- Michael Kozlow, Ph.D., Education Quality and Accountability Office (Ontario)
- Scott Marion, Ph.D., *Center for Assessment*
- Stanley N. Rabinowitz, Ph.D., *WestED*
- Ed Wiley, Ph.D., *University of Colorado*

Questions